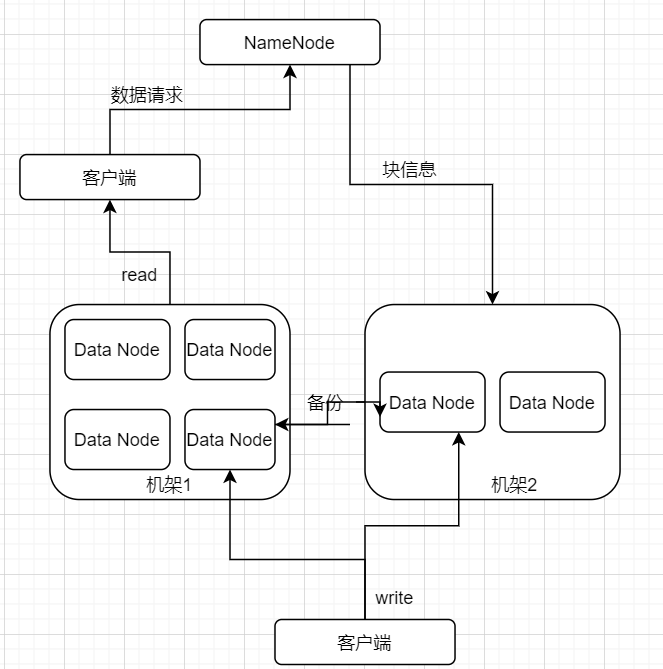
# 一. Hadoop 理论

## 1. Hadoop 体系结构

HDFS 和MapReduce 是Hadoop 的两大核心，整个Hadoop的体系结构主要是通过HDFS来实现分布式存储的底层支持的，通过MapReduce来实现并行任务处理的程序支持。

HDFS 的体系结构：

HDFS采用了主从（Master/Slave）结构模型，一个HDFS 集群是由一个NameNode和若干个DataNode组成的。其中NameNode作为主服务器管理文件系统的命名空间和客户端对文件的访问操作；集群中的DataNode管理存储的数据，HDFS允许用户以文件的形式存储数据。即文件被分成若干个数据块，而且这若干数据块存放在一组DataNode上，NameNode执行文件系统的命名空间操作，包括文件或目录的打开、关闭、重命名，并在NameNode的统一调度下进行数据块的创建、删除和复制空间。



MapReduce的体系架构：

MapReduce是一种并行编程模式，开发人员利用这种并行编程模式编写出分布式并行程序。MapReduce是由一个单独运行在主节点的JobTracker和运行在每个集群从节点的TaskTracker共同组成的。主节点负责调度构成一个作业的所有任务，这些任务分布在不同的从节点上，主节点监控它们的执行情况，并且重新执行之前失败的任务；从节点仅负责由主节点指派的任务，当一个Job被提交时，JobTracker接收到提交作业和其配置信息之后，就会将配置信息等分发给从节点，同时调度任务并监控TaskTracker的执行。

Hadoop 计算模型 – MapReduce

一个Map/Reduce作业（Job）通常会把输入的数据集切分成若干独立的数据块，由Map任务（Task）以完全并行的方式处理它们。框架首先对Map的输出进行排序，然后把结果输入给Reduce任务。通常作业的输入和输出都会被存储在文件系统中。

Map/Reduce框架由一个单独的Master JobTracker和集群节点上的Slave TaskTracker 共同组成。Master负责调度一个作业的所有任务，而Slave及负责执行由Master指派的任务。

# 2. Hadoop数据管理

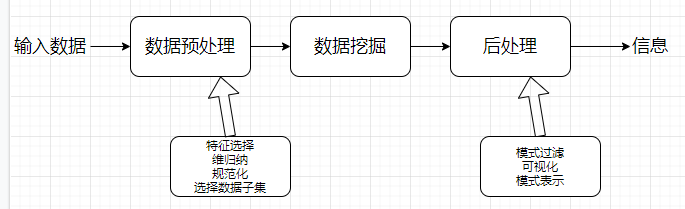
Hadoop 的数据管理主要包括Hadoop的分布式文件系统HDFS、分布式数据库HBase和数据仓库工具Hive。

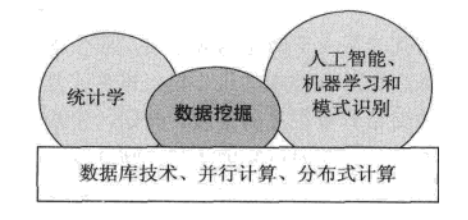
HDFS是分布式计算的存储基石，主要有三个角色来进行文件系统的管理，NameNode、DataNode和Client，NameNode是文件系统的管理者，主要负责管理文件系统的命名空间、集群配置信息和存储块的复制，NameNode会将文件系统的Metadata存储在内存中，这些信息主要包括文件信息、每一个文件对应的文件块的信息和每一个文件块在DataNode中的信息等。DataNode是文件存储的基本单元，它将文件块（Block）存储在本地文件系统中，保存了所有Block的Metadata，同时周期性地将所有存在的Block信息发送给NameNode。Client就是需要获取分布式文件系统文件的应用程序。

# 数据挖掘

## 什么是数据挖掘

数据挖掘技术用于侦察大型数据库，发现先前未知的有用模式，可以观测未来观测结果。是数据库中知识发现（KDD，knowledge discovery in database）不可缺少的一部分，而KDD是将未加工的数据转换为有用信息的过程。

KDD 过程



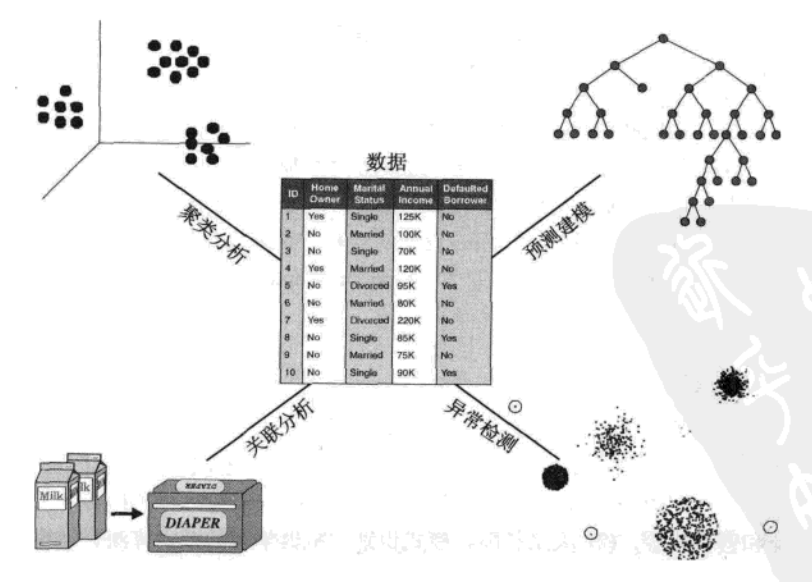
数据挖掘与其他领域之间的联系

## 数据挖掘的任务

数据挖掘任务分为以下两大类。

**预测任务**：根据其他属性的值，预测有特定属性的值。被预测的属性一般呈目标变量或因变量，而用来做预测的属性称说明变量或自变量。

**描述任务：**目标是导出概括数据中潜在联系的模式（相关、趋势、聚类、轨迹和异常）。描述性数据挖掘任务通常侦查性的，并且常常需要后处理技术验证和解释结果。



其余四种主要数据挖掘任务

**预测建模（predictive modeling）：**涉及以说明变量函数的方式为目标变量建立模型。有两类预测建模任务：分类（classification）用于预测离散的目标变量；回归（regression）用于预测连续的目标变量。

**关联分析（association analysis）:** 用来发现描述数据中强关联特征的模式。

**聚类分析（cluster analysis）:** 用来发现紧密相关的观测值组群，使得与属于不同簇的观测值相比，属于同一簇的观测值相互之间尽可能类似。聚类可用来对相关的顾客分组、找出显著影响地球气候的海洋区域以及压缩数据等。

异常检测（anomaly detection）：识别其特征显著不同于其他数据的观测值。这样的观测值称为异常点（anomaly）或离群点（outlier）。异常检测算法的目标是发现真正的异常点，而避免错误地将正常的对象标注为异常点。好的异常检测器必须具有高检测率和低报错率。异常检测的应用包括检测欺诈、网络攻击、疾病的不寻常模式、生态系统扰动等。

## 数据

### 数据类型

数据集看作是数据对象的集合。数据对象有时也叫做记录、点、向量、模式、事件、案例、样本、观测或实体。数据对象用一组刻画对象基本特性（如物体质量或事件发生时间）的属性描述。属性有时也叫做变量、特性、字段、特征或者维。

* + 1. 属性与度量

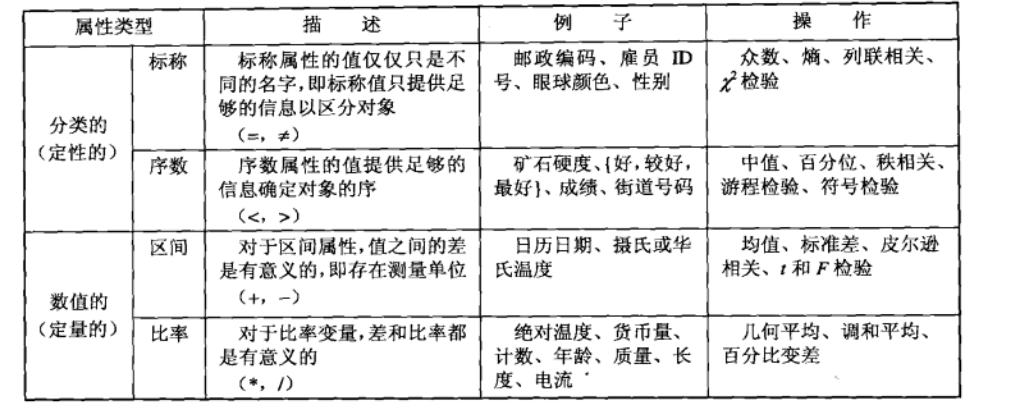
1. 属性

属性(attribute)是对象的性质或特性，因对象而已，或随时间而变化。

测量标度（measurement scale）是将数值或符号值与对象的属性相关联的规则（函数）。

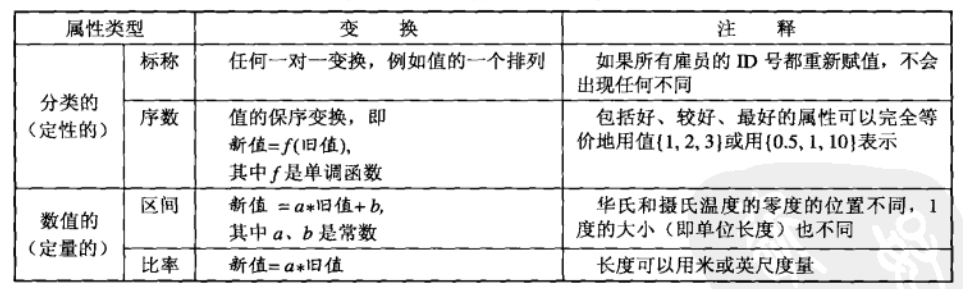
1. 属性类型

属性类型包含四种：标称（nominal）、序数（ordinal）、区间（interval）、比率（ratio）。



不同的属性类型

也可用允许的变换来定义属性类型。



定义属性层次的变换

1. 用值的个数描述属性

区分属性的一种独立方法是根据属性可能取值的个数来判断。

**离散的**：离散属性具有有限个值或无限个可数个值。这样的属性是可以分类的。二元属性（binary attribute）是离散属性的一种特殊情况，只接受两个值，如真/假、是/否、男/女或0/1。

**连续的：**连续属性是取实数值的属性。通常连续属性用浮点变量表示。

### 数据集的类型

1. 数据集的一般特性：

**维度**：数据集的维度是数据集中的对象具有的属性数目。分析高纬度数据有时会陷入所谓维灾难（curse of dimensionality）。数据预处理的一个重要动机就是减少维度，称为维归纳（dimensionality reduction）。

**稀疏性**（sparsity）：如具有非对称特征的数据集，一个对象的大部分属性上都是为0；在许多情况下，非零项还不到1%。

**分辨率**：不同分辨率下的数据性质也不同。

1. 记录数据

记录数据的不同变体：

**事务数据或购物篮数据**：事务数据（transaction data）：是一种特殊类型的记录数据，其中每个记录（事务）涉及一系列的项。

**数据矩阵（模式矩阵）**：数据对象集可以用一个mxn矩阵表示，m行，一个对象一行；n列，一个属性一个列。

稀疏数据矩阵：数据矩阵的一种特殊情况，其中属性的类型相同并且是非对称的，即只有非零值才是重要的。文档可以用词向量表示，每一个词是向量的一个分量（属性），而每个分量的值是对应词在文档出现的次数。文档集合的这种表示通常称作文档-词矩阵。

1. 基于图形的数据

图形可以方便有效地表示数据，一种是带有对象之间联系的数据。另一种是具有图形对象的数据。

1. 有序数据

某些数据类型，属性具有涉及时间或空间序的联系。

**时序数据（sequential data）：**时序数据也成为时间数据，可以看作记录数据的扩充，其中每个记录包含一个与之相关联的时间。

**序列数据（sequence data**）是一个数据集合，是各个实体的序列。没有时间戳，只是有序序列考虑向的位置。

**时间序列数据（time series data）：**是一种特殊的时序数据，记录每个记录都是一个时间序列，即一段时间以来的测量序列。在分析时间数据时，最重要的是考虑时间自相关。

**空间数据：**有些对象还具有空间属性，空间数据的一个重要特点是空间自相关性，即物理上靠近的对象趋于在其他方面也相似。

1. 数据质量
2. 测量和数据收集问题

**测量误差**：指测量过程中导致的问题。对于连续属性，测量值与实际值的差称为误差（error）。

**数据收集错误：**是指诸如一楼数据对象或属性值，或不当地包含其他数据对象等错误。

1. 噪声和伪像

**噪声**是测量误差的随机部分。

数据错误可能是更确定性现象的结果，如一组照片在同一地方出现条纹。数据的这种确定性失真常称为伪像。

1. 精度、偏倚和准确率

测量过程和结果数据的质量用精度和偏倚度量。

**精度（precision）:** （同一个量的）重复测量值之间的接近程度。

**偏倚（bias）：**测量值与被测量之间的系统的变差。

**准确率（accuracy）:** 被测量的测量值与实际值之间的接近度。准确率的一个重要方面是有效数字（significant digit）的使用。其目标是仅使用数据精度所能确定的数字位数表示测量或计算结果。

1. 离群点

离群点是在某种意义上具有不同于数据集中其他大部分数据对象的特征的数据对象，或是相对于改属性的典型值来说不训成的属性值，也称为异常（anomalous）对象或异常值。

1. 遗漏值

一个对象遗漏一个或多个属性值的情况并不少见。处理遗漏值的策略如下：

**删除数据对象或属性：**一个简单而有效的策略是删除具有遗漏值的数据对象。

**估计遗漏值：**遗漏值可以使用其他值来估计（插值）。

**在分析时忽略遗漏值**

6）不一致的值：对数据进行纠正。

7）重复数据：去重操作。

### 数据预处理

使用特征（feature）或变量（variable）指代属性（attribute）。

#### 聚集

聚集就是将两个或多个对象合并成单个对象。

定量属性（如价格）通常通过求和或求平均值进行聚集。定性属性（如商品）可以忽略或汇总成在一个商店销量的所有商品的集合。

聚集的优点：数据归纳导致的较小数据集需要较少的内存和处理时间，通过高层而不是低层数据视图，聚集起到了范围或标度转换的作用。聚集的缺点可能会丢失一些细节。

#### 抽样

抽样是一种选择数据对象子集进行分析的常用方法，抽样具有代表性，前提是他近似地具有与原数据相同的（感兴趣的）性质，

1. 抽样方法：

**简单随机抽样（simple random sampling）:** 包含无放回抽样（每个选中项立即从构成总体的所有对象集中删除）和有放回抽样（对象被选中时不从总体中删除）。

**分层抽样（stratified sampling）:** 当总体有不同类型的对象组成，每种类型的对象数量差别很大时，简单随机抽样不能充分代表不太频繁出现的对象类型。这时需要分层抽样，从预先指定的组开始抽样。

1. 渐进抽样

合适的样本容量可能很难确定，因此需要使用自适应（adaptive）或渐进抽样（progressive sampling）方法。

#### 维归约

数据集可能包含大量特征。比如一个文档的集合，每个文档都是一个向量，其分量是文档中出现的每个词的频率。

维归约的好处是如果维度（数据属性的个数）较低，许多数据挖掘算法的效果就会更好，当模型只涉及较少的属性时，模型更容易理解，还可以更容易实现数据可视化。

维归约通常用于这样技术：通过创建新属性，将一些旧属性合并在一起来降低数据集的维度。通过选择旧属性的子集得到新属性。这种维归约称为特征子集选择或特征选择。

1. **维灾难**

维灾难是指这样一种现象：随着数据维度的增加，许多数据分析变得非常困难。数据在它所占据的空间中越来越稀疏，对于分类，可能意味着没有足够的数据对象来创建模型，将所有可能的对象可靠地指派到一个类。对于聚类，点之间的密度和距离的定义失去了意义，结果是，对于高维数据，许多分类和聚类算法都存在问题 ——— 分类准确率降低，聚类质量下降。

1. **维归约的线性代数技术**

维归约最常用的方法是使用线性代数技术，将数据由高维空间投影到低维空间，特别是对于连续数据。**主成成分分析（Principal components Analysis, PCA）**是一种用于连续属性的线性代数技术，它找出新的属性（主成分），这些属性石原属性的线性组合，是相互正交的，并且捕获了最大变差。**奇异值分解（Singular Value Decomposition, SVD）**是一种线性代数技术，与PCA相关，并且用于维归约。

#### 特征子集选择

降低维度的另一种方法是仅使用特征的一个子集。冗余特征重复了包含一个或多个其他属性中的许多或所有信息。不相关特征包含对于手头的数据挖掘任务几乎完全没用的信息。

特征选择的理想方法就是将所有可能的特征子集作为感兴趣的数据挖掘算法的输入，然后选取产生最好结果的子集。这种方法的优点是反映了最终使用的数据挖掘算法的目的和偏好。有三种标准的特征选择方法：嵌入、过滤和包装。

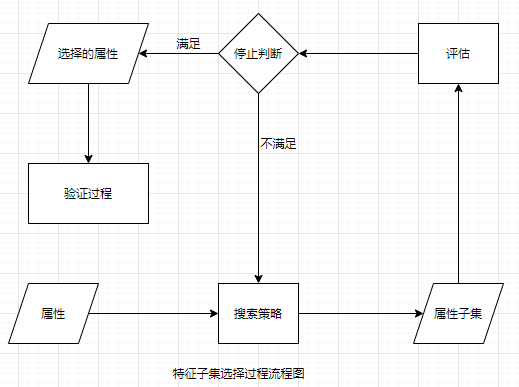
**嵌入方法（embedded approach）**：决定使用哪些属性和忽略哪些属性。

**过滤方法（filter approach）**:使用某种独立于数据挖掘任务的方法，进行特征选择，选择属性的集合，属性对之间的相关度尽可能低。

**包装方法（wrapper approach）：**将目标数据挖掘算法作为黑盒，通常不枚举所有可能的子集来找出最佳属性子集。

1. 特征子集选择体系结构：

特征选择过程可以看作由四部分组成：子集评估度量、控制新的特征子集产生的搜索策略、停止搜索判断和验证过程。过滤方法和包装方法唯一不同之处在于它们使用了不同的特征子集评估方法。包装方法的子集评估使用目标数据挖掘算法，过滤方法子集评估技术不同于目标挖掘算法。



1. 特征加权

特征加权是一种保留或删除特征的方法。特征越重要，所赋予的权值越大，而不太重要放入特征赋予较小的权值。

#### 特征创建

创建新属性的相关方法：特征提取、映射数据到新的空间和特征构造。

1. **特征提取**

由原数据创建新的特征集称作特征提取。

1. **映射数据到新的空间**

使用一种完全不同的视角挖掘数据可能揭示出重要和有趣的特征。对时间序列数据来说，常常包含周期模式，若噪声大，很难检测到这些模式，需要对时间序列实**施傅里叶变换**（Fourier transform，转换成频率信息明显的表示。除此之外还有小波变换（wavelet transform）方法

1. **特征构造**

有时原始数据集的特征具有必要的信息，但其形式不适合数据挖掘算法，因此，一个或多个由原特征构造的新特征可能比原特征更有用。

#### 离散化和二元化

将连续属性转换成分类属性叫做离散化（discretization），连续和离散属性可能都需要变成一个或多个二元属性叫做二元化（binarization）。

1. 二元化

对于m个分类值，将每个原始值唯一地赋予区间[0, m-1]中的一个整数，将这些整数都意义转化成一个二进制数。

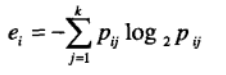
1. 连续属性离散化

离散化问题就是决定选择多少个分割点和确定分割点位置的问题。

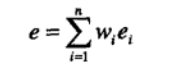
**非监督离散化**：用于分类的离散化方法之间的根本区别在于使用类信息（监督，supervised）还是不使用类信息（非监督，unsupervised）。非监督离散化的方法有等宽、等频率（等深）、K均值方法。

监督离散化：基于熵的方法是最有前途的离散化方法之一。

定义熵：设K是不同的类标号数，mi 是某划分的第i个区间中值的个数，而mij 是区间i中类j的值的个数。第i个区间的熵ei 由如下等式给出



其中，pij = mij /mi 是第i个区间中类j的概率（值的比例）。该划分的总熵e是每个区间的熵的加权平均，即



其中m是值的个数，wi = mi / m是第i个区间的值的比例，而n是区间个数。直观上，区间的熵是区间纯度的度量。

#### 变量变换

变量变换是指用于变量的所有值的变换。

**简单函数**

**标准化（standardization）或规范化（normalization）**是使整个值的集合具有特定的性质。

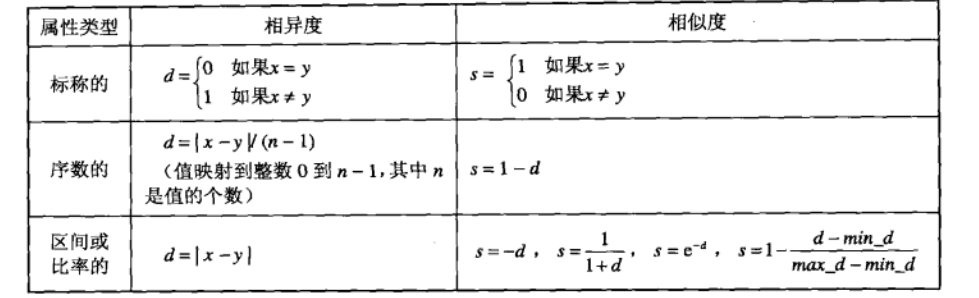
### 相似性和相异性的度量

1. **相似度和相异度**

两个对象之间的相似度（similarity）就是这连个对象相似程度数值度量。相似度是非负的，并常常在0（不相似）和1（完全相似）之间取值。

两个对象之间的相异度（dissimilarity）是这两个对象差异程度的数值度量。对象越类似，它们的相异度就越低。距离（distance）用作相异度的同义词，相异度在【0,1】中取值。

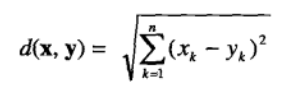
1. **简单属性之间的相似度和相异度**

简单属性的相似度和相异度

1. **数据对象之间的相异度**

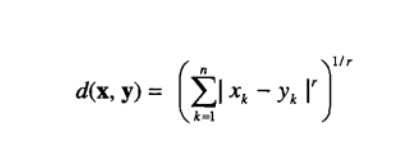
**距离**

欧几里得距离（Euclidean distance）公式：



n是维数，而xk和yk分别是x和y的第k个属性值（分量）。点与点集合、点与点之间的距离表格称为**欧几里得矩阵**（distance matrix）。

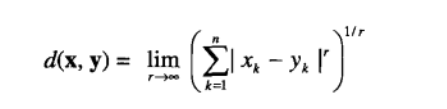
**闵可夫斯基距离（Minkowski distance）:**



r是参数

**1）**r=1，城市街区（也成曼哈顿、出租车、L1范数）距离。汉明距离（Hamming distance）,它是两个具有二元属性的对象（即两个二元向量）之间不同的二进制个数。

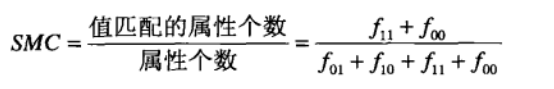
**2）**r=2，欧几里得距离（L2范数）。

**3)** r = IMG_256∞， 上确界（Lmax或L∞范数）距离。这是对象属性之间的最大距离。L∞距离公式：

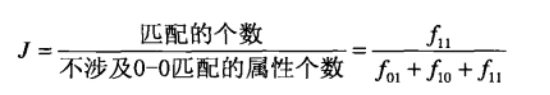
1. **邻近性度量的例子**
2. **二元数据的相似性度量**

两个仅包含二元属性的对象之间的相似性度也称为相似系数（similarity coefficient），并且通常在0到1之间取值，1代表两个对象完全相似，0代表对象一点也不相似。

**简单匹配系数（simple matching coeficient， SMC）**定义如下：

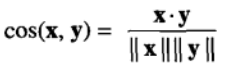


Jaccard系数（Jaccard Coefficient）来处理仅包含非对称的二元属性的对象，通常用符号J表示：



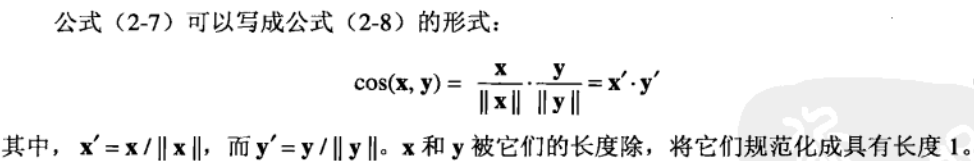
1. **余弦相似度**

文档用向量表示，响亮的每个属性代表一个特定的词在文档中出现的频率。文档相似度量不仅应当像Jaccard度量一样需要忽略0-0匹配，而且还必须能够处理非二元向量。最常用的文档相似性度量方法是余弦相似度（cosine similarity）。x，y是两个文档向量，



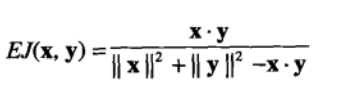


以上公式可以写成以下公式形式：



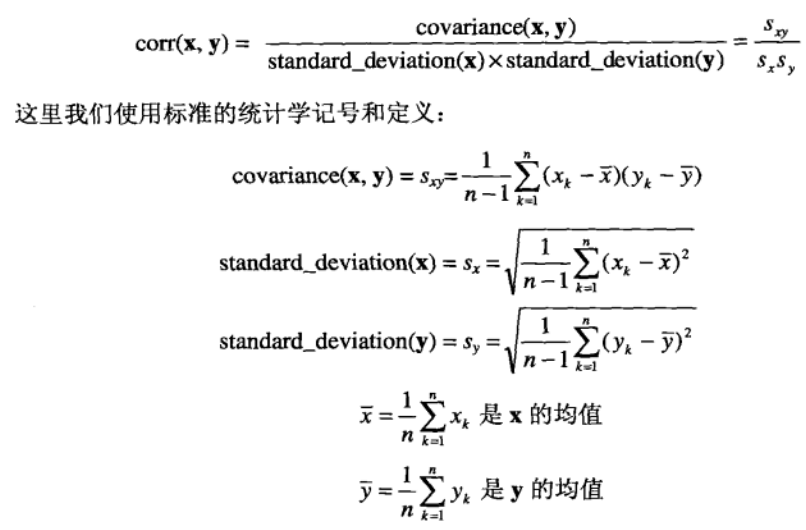
1. **广义Jaccard系数（Tanimoto系数）**

该系数用EJ表示，有以下公式定义：



1. **相关性**

两个数据对象x 和y 之间的皮尔森相关（Pearson’s correlation）系数由以下公式定义：



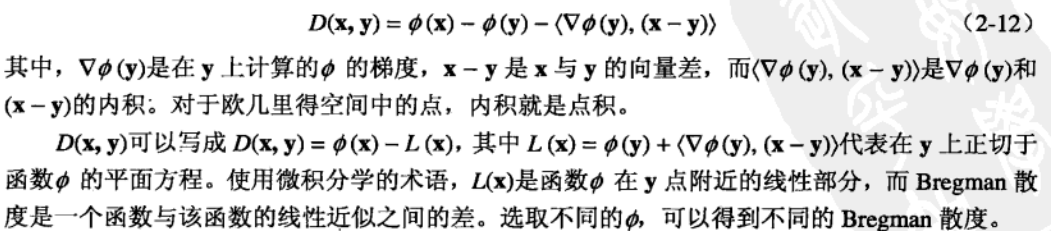
**完全相关**：相关度总是在-1到1之间取值。相关度为1（-1）意味着x和y具有完全正（负）线性关系，即Xk = ayk+b，其中a和b是常数。

**非线性关系：**如果相关度为0，则两个数据对象的属性之间不存在线性关系，也可能存在非线性关系。

**相关性可视化：**通过绘制对应属性值对可以很容易判定两个数据对象x和y之间的相关性。

**Bregman散度**是损失或失真函数（损失函数的目的是度量用x近似y导致的失真或损失）。

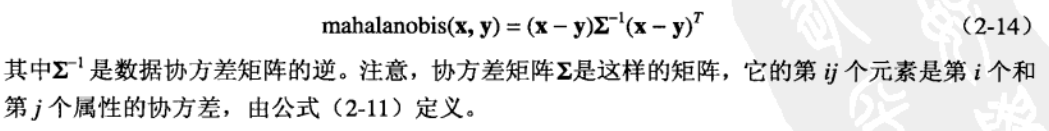
定义如下：给定一个严格凸函数∅（连同一些通常满足的适度限制），由该函数生成的Bregman散度（损失函数）D(x,y)通过下面的公式给出：



1. **邻近性计算问题**

**1）距离度量的标准化和相关性**

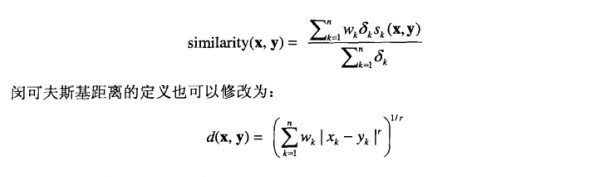
距离度量的一个重要问题是当属性具有不同的值域时如何处理。两个对象（向量）x和y之间的Mahalanobis距离定义为：



1. 组合异种属性的相似度

当属性具有不同类型时，最直接的方法是计算每个属性之间的相似度，然后使用一种导致0和1之间相似度的方法组合这些相似度。总相似度一般定义为所有属性相似度的平均值。

1. 使用权值



## 探索数据

* 1. 汇总统计

汇总统计是量化的（如均值和标准差），用单个数或数的小集合捕获可能很大的值集的各种特征。

1. 频率和众数

给定一个无序、分类的值的集合，为了进一步刻画值的性质，除计算特定数据集中每个值出现的频率外没有多少的事情可做。给定一个在{v1, v2 ,v3，………, Vk}上取值的分类属性x和m个对象的集合，值Vi的频率定义为：



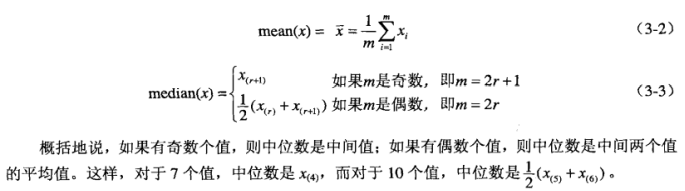
分类属性的众数（mode）是具有最高频率的值。

1. 百分位数

对于有序数据，考虑值集的百分位数更有意义。

1. 位置度量：举止和中位数

对于连续数据，两个使用最广泛的汇总统计是均值和中位数。



截断均值概念，指定0和100之间的百分位数p,丢弃高端和低端（p/2）%的数据，然后用常会的方法计算均值，所得结果即为截断均值。

1. 散步度量：极差和方差

极差(range)定义：range(x) = max(x)-min(x)= x(m) – x(l)

方差（variance）：通常属性x的观测值的方差记作Sx­­2­，标准差（standard deviation）是方差的平方根，记作Sx,它与x具有相同的单位。



均值可能被离群值扭曲，并且由于方差用均值计算，因此它对离群值敏感。

绝对平均偏差（absolute average deviation, ADD）：



中位数绝对偏差（median absolute deviation, MAD）:



四分位数极差（interquartile range, IQR）

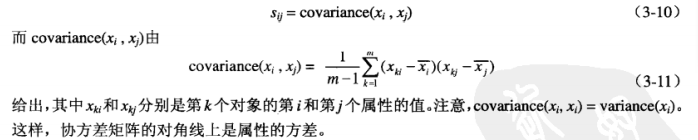


5) 多元汇总统计

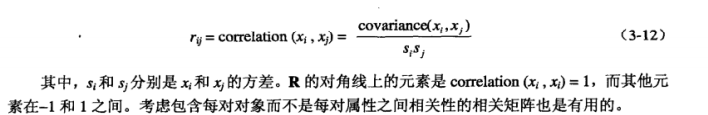
包含多个属性的数据（多元数据）的位置度量可以通过分别计算每个属性的均值或中位数得到，这样，给定一个数据集，数据对象的均值：



对于具有联系变量的数据，数据的散步更多地用协方差矩阵（convariance matrix）S 表示，其中S的第ij个元素Sij始数据的第i个和第j个属性的协方差。这样的，如果Xi和xj分别是第i个和第j个属性，则



协方差的值接近0表明两个变量不具有线性关系，但是不能仅靠协方差的值来确定两个变量之间的关联程度。因为两个属性的相关性直接指出两个属性线性相关的程度，对于数据探索，相关协比协方差更可取。相关矩阵（correlation matrix）R 的第ij元素是数据的第i个和第j个属性之间的相关性。如果xi和xj分别是第i个和第j个属性，则：



6）汇总数据的其他方法